DOCUMENT RESUME

ED 111 837 TM 004 804

AUTHOR Fishbein, Ronald L.
TITLE An Investigation of the Fairness of the Items of a Test Battery.
PUB DATE [Apr 75]
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2. 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Ethnic Groups; Grade 8; *Item Analysis; Mathematics; *Norm Referenced Tests; Occupational Tests; Reading Tests; Sex Discrimination; *Statistical Analysis; *Test Bias

ABSTRACT

This study develops a procedure for detecting items which are biased for particular ethnic groups and utilizes this procedure to evaluate the fairness of reading, mathematics, and occupational information test items for several ethnic groups. The population for each ethnic group was chosen from examinees administered the 1973 version of the Florida Eighth Grade Testing Program (FEGTP). In this study, an item was considered biased if it manifested an Item X Group interaction. Few biased items were detected on the Reading, Mathematics, and Occupational Information tests of the FEGTP. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED111837

ABSTRACT*

The purpose of this study was to develop a procedure for detecting items which are biased for particular ethnic groups and to utilize this procedure to evaluate the fairness of reading, mathematics, and occupational information test items for several ethnic groups.

The population for each ethnic group was chosen from examinees administered the 1973 version of the Florida Eighth Grade Testing Program (FEGTP).

In this study, an item was considered biased if it manifested an Item X Group interaction. Few biased items were detected on the Reading, Mathematics, and Occupational Information tests of the FEGTP.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM004 804

---

*Fishbein, R. L. An investigation of the fairness of the items of a test battery. Paper presented at the meeting of the National Council on Measurement in Education, Washington D.C., March 1975.

AN INVESTIGATION OF THE FAIRNESS OF THE ITEMS OF A TEST BATTERY[1]

Ronald L. Fishbein

Michigan Department of Education

This study focused on assessing the fairness of the items of several tests when a predicted criterion variable was unavailable. A procedure was developed for detecting items which are unfair or biased for particular ethnic groups, and this procedure was utilized to evaluate the fairness of reading, mathematics, and occupational information test items for several ethnic groups.

The definition of bias used in this study should not necessarily be equated with the term "cultural bias." For the purposes of this study an item was considered biased against a group compared with another group if the item manifested an Item X Group interaction. The statistical procedure employed to detect interaction identified items where a group's mean on an item was higher or lower than another group's mean on the item by an amount higher or lower than would be expected from a comparison of both groups' total test performance (see Cleary & Hilton, 1968).

The bias of the Reading (vocabulary and comprehension), Mathematics (computation and problem solving), and Occupational Information test items of the 1973 Florida Eighth Grade Testing Program (FEGTP) was assessed in this study. The FEGTP is a basic skills test battery administered annually to virtually every eighth grade student in the state of Florida. The three tests evaluated were

norm referenced.

The groups considered in this study were: (1) White Caucasians, (2) Black Afro-Americans, (3) American Indians, (4) Orientals, (5) Puerto Rican Americans, (6) Cuban Americans, (7) Males, (8) Females, (9) Urban examinees, and (10) Rural examinees. Examinee classifications were determined from the information provided by the examinee on the answer sheet of the test battery under the categories Race Code and Sex. In addition, an examinee was classified as urban if he participated in the 1973 test administration in a county with at least 96.1% of the population defined as urban for 1970 by the U. S. Department of Commerce, Bureau of the Census, and as rural if he participated in a county with 0.0% of the population defined as urban.

Several previous studies have evaluated test fairness without the use of a predicted criterion variable. These studies have used analysis of variance (ANOVA) procedures to detect significant Item X Group interactions. For example, Cardall and Coffman (1964) assessed the fairness of the items of the Scholastic Aptitude Test (SAT) for Rural, Urban, and Black examinees using a two factor ANOVA design with repeated measures on items. The significant Item X Group interactions indicated that some items of the SAT may have had different relative difficulties for the groups examined. Similar investigations (Angoff & Sharon, 1974; Cleary & Hilton, 1968) have also detected significant Item X Group interactions.

Angoff and Sharon (1974) noted that a major limitation of

using ANOVA to detect Item X Group interaction is the failure to detect the specific items contributing to the interaction. They attempted to overcome this shortcoming by producing a bivariate plot of item difficulty values for each pair of groups being compared and calculated the perpendicular distance of each point from the major axis of the elliptical plot of item points. However, Angoff and Sharon did not attempt to specify how deviant an item had to be before it should be labelled as biased. The technique used in this study detected significant Item X Group interactions at the level of the individual item.

METHOD

Samples

Samples were chosen from the data of the 1973 administration of the FEGTP. Five systematic samples of 225 examinees in each sample were chosen from the population of White examinees; five systematic samples of 225 examinees in each sample were chosen from the population of Black examinees; and five systematic samples of 225 examinees in each sample were chosen from the population of Cuban American examinees. The samples from the White, Black, and Cuban American populations were mutually exclusive. Systematic samples of 225 examinees were also chosen from the populations of each of the following groups: American Indians, Puerto Rican Americans, Males, Females, Urban examinees, and Rural examinees. In addition, a systematic sample of 224 examinees was chosen from the population of Oriental examinees.[2]

All samples from each population were chosen without replacement. Since the item responses of the examinees who participated in the testing program were grouped by county, and by school within each county, the systematic samples chosen were, for practical purposes, equivalent to stratified samples where the number of examinees chosen from any school represented approximately the proportion of examinees from that school in the population of the group sampled.

Procedure

In this study the term bias was always used in a comparative sense. An item was biased against a group compared with another group. For the purpose of determining whether test items are biased against certain ethnic groups, an item was considered biased against a group compared with another group if the item manifested an Item X Group interaction. Stated differently, an item was considered biased if the difference in performance on the item for the two groups was significantly different than the difference in their overall performance on the test. If the difference in performance on an item was significantly less than the overall difference in performance between the two samples of a comparison, then the item was considered biased against the group having the higher overall performance. If the difference in performance on an item was significantly greater than the overall difference in performance between the two groups of a comparison, then the item was considered biased against the group having the lower overall performance. The 29 comparisons made on each test that was assessed for bias are shown in Table 1.

- - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - -

The following procedure was employed to determine whether the reading, mathematics, and occupational information test items of the FEGTP were biased against certain ethnic groups.

The population p value on the ith test for the jth group on the kth item was set equal to p_{ijk}, and the average population p value on the ith test for the jth group was set equal to $\bar{p}_{ij}$. Then, for example, to test the fairness of the third item on the second test for groups one and six the following statistical hypotheses were formed, where $\Delta = \bar{p}_{21} - \bar{p}_{26}$.

$$H_0: \quad p_{213} - p_{263} = \Delta \tag{1}$$

$$H_1: \quad p_{213} - p_{263} \neq \Delta \tag{2}$$

The null hypothesis was tested by forming the following confidence interval, where p_{ijk} equaled the p value on the ith test for the jth group on the kth item, and N_{ij} equaled the number of examinees from the jth group who had taken the ith test (see Marascuilo, 1971). If the confidence interval did not include Δ, then the null hypothesis was rejected, with the probability of a Type I error equal

$$p_{213} - p_{263} = (\hat{p}_{213} - \hat{p}_{263}) \pm Z_{\alpha/2} \sqrt{\frac{(\hat{p}_{213})(1-\hat{p}_{213})}{N_{21}} + \frac{(\hat{p}_{263})(1-\hat{p}_{263})}{N_{26}}} \tag{3}$$

to α.

Each of the three tests assessed for bias was considered an experiment. For each item of each test that was assessed for bias, the 29 comparisons listed in Table 1 were made with $\alpha = .0002$ for each comparison. With α equal to .0002 for a comparison, the Type I error rate for an item was approximately .0058. Therefore, the test α for reading was approximately .348; the test α for mathematics was approximately .452; and the test α for occupational information was approximately .232. If one used an alternative hypothesis that the least significant difference of interest for Equation 2 was .25, then the power of each statistical comparison was approximately .94 (see Marascuilo, 1971, p. 301). This calculation assumed a maximum standard error of the difference between the two proportions. Cohen (1969) has defined a difference between two independent proportions of approximately .25 as a medium effect size.

RESULTS

The major finding of this study was that there were few biased items on the Reading, Mathematics, and Occupational Information tests of the FEGTP when an item was defined as biased if it manifested an Item X Group interaction. The percentage of biased comparisons on the Reading test was 3.68; the percentage of biased comparisons on the Mathematics test was 1.89; and the percentage of biased comparisons on the Occupational Information test was 1.58. However, it should be pointed out that 7 out of 1,740 comparisons on the Reading test, 86 out of 2,262 comparisons on the Mathematics test, and 19

out of 1,160 comparisons on the Occupational Information test were eliminated from consideration because of a ceiling effect--the p values for both groups were high enough that it was impossible, or inconceivable, for the group with the larger test mean to outscore the comparison group by a value as large as Δ. Also, 19 comparisons were eliminated from the Reading test because of a floor effect--the p values for both groups were below, at, or slightly above the chance level.

Table 2 indicates the percentage of biased items on the Reading test for each comparison of the study. The Cuban-Indian comparison,

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

which contained the highest percentage of biased items on the Reading test, had an equal number of items biased against Cubans as American Indians. On the White-Cuban comparison there were 11 instances of bias against Cubans and 9 instances of bias against Whites. However, all nine instances of bias against Whites on the White-Cuban comparison were vocabulary items which resembled the Spanish translation and, therefore, gave an unusual advantage to Cuban American examinees. On the Black-Cuban comparison there were 7 instances of bias against Cubans and 13 instances of bias against Blacks. On the White-Black comparison there were seven instances of bias against Blacks and six instances of bias against Whites. There was no evidence of bias on

the remaining Reading test comparisons.

Table 3 indicates the percentage of biased items on the Mathematics test for each comparison of the study. The White-Indian comparison, which had the highest percentage of biased items on the

- - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - -

Mathematics test, contained 11 items biased against American Indians and no items biased against Whites. On the Oriental-Indian comparison there was one instance of bias against Orientals and four instances of bias against American Indians. On the Black-Oriental comparison there were three instances of bias against Blacks and no instances of bias against Orientals. There was no evidence of bias on the remaining Mathematics test comparisons.

Table 4 indicates the percentage of biased items on the Occupational Information test for each comparison of the study. The only comparisons which showed bias on the Occupational Information

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

test were White-Black, Black-Oriental, Black-Cuban, and Male-Female. The Black-Cuban comparison showed the highest percentage of biased items on the Occupational Information test. There were five instances of bias against Blacks and four instances of bias against Cubans.

Specifically, the following generalizations concerning the Reading, Mathematics, and Occupational Information tests seem warranted. There was a greater tendency for reading vocabulary items to exhibit bias than reading comprehension items. This was true even after taking into account that the Vocabulary subtest was twice as long as the Comprehension subtest. This tendency was especially pronounced for Blacks, where all 20 instances of bias against Blacks on the Reading test were vocabulary items. The researcher was unable to explain this unexpected occurrence.

There was virtually no evidence of bias on Male-Female and Urban-Rural comparisons. There were also few instances of bias against Oriental and Puerto Rican examinees. When bias was detected, items were most often biased against Whites, Blacks, Cubans, and American Indians. Bias against Blacks, Cubans, and Indians was expected, but bias against Whites was unforeseen. However, items biased against Whites were often detected on White-Cuban comparisons and, as mentioned previously, could be explained because the biased item was a vocabulary word which resembled the Spanish translation.

There was a tendency for a relatively higher percentage of comparisons to show bias on the Problem Solving section of the Mathematics test than on the Computation section. This result was consistent with expectations, unlike the finding that a higher percentage of reading vocabulary items was biased than reading comprehension items.

DISCUSSION AND RECOMMENDATIONS

This study failed to detect a substantial degree of Item X

Group interaction for the items of the Reading, Mathematics, and Occupational Information tests of the FEGTP. Since the test items which were assessed for bias are representative of basic skills test items given below the college level, and to the extent that Florida students are representative of the nation, similar results might be obtained with other achievement batteries in other areas of the country. Assuming that similar results would be obtained, what would be demonstrated? The logic of statistical inference does not permit one to prove a null hypothesis, and one would only be justified in saying that since the hypothesis of no Item X Group interaction was usually not rejected, one can continue to entertain the hypothesis that there is little interaction. Even if the researcher could have proved the null hypothesis for every comparison of this study, item fairness would not have been demonstrated. A finding of no Item X Group interaction means that a test item is functioning in a homogeneous manner in terms of the relative difficulty for the groups of a comparison. However, the possibility exists that all of the items of a test may be biased against a particular group, but none of the items would display interaction because they are all biased in a similar manner. It is also possible that an item detected as biased was actually fair, but was labeled biased because most of the other items on the test were biased.

The reading and mathematics test items assessed in this study were developed by a major commercial testing company and the occupational information test items were developed by the staff of the FEGTP. These items had undergone considerable editing and all three

tests had been administered to representative Florida samples before the final versions were printed. Even with these safeguards, possible flaws were found in many of the biased items. Once weaknesses were detected, logical revisions seemed possible.

It would be extremely advantageous if biased items would be detected during the test development stage. This would be an especially important consideration if the degree of bias on other test batteries were found to be greater than that of the FEGTP.

Several of the items which manifested bias were examined by a Black and several native Spanish speaking graduate students. They were able to propose logical explanations for the behavior of many biased items. The present writer, a White male, was unable to detect many of these weaknesses. This would indicate that test fairness would probably be improved if members of minority groups would edit items on standardized tests. It also seems logical that test fairness would be improved if minority groups were included on committees which determine the objectives and content to be tested. This seems particularly important in the development of criterion referenced tests.

In conclusion, it should be emphasized that a question as controversial as the fairness of psychological test items cannot be resolved by psychometric debate. When minority group is no longer synonomous with lower scoring group, the issue of test bias will cease to exist.

REFERENCES

Angoff, W. H., & Sharon, A. T. The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 1974, 34, 807-816.

Cardall, C., & Cofman, W. E. A method for comparing the performance of different groups on the items in a test. Research Bulletin-64-61. Princeton: Educational Testing Service, 1964.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Marascuilo, L. A. Statistical methods for behavioral science research. New York: McGraw-Hill, 1971.

FOOTNOTES

[1]This paper is based upon the author's Ph.D. dissertation submitted to the faculty of the Educational Evaluation and Research Design Program, The Florida State University. Helpful suggestions were made by Jacob G. Beard (major professor), Harman D. Burck, Garrett R. Foster, John R. Hills, Howard W. Stoker, and Gerald J. Schluck.

The computer programs for this study were written by Dr. Philippe Olivier and Mrs. Marjorie Olivier.

[2]The sample of Oriental examinees was 224 because of a programming error.

TABLE 1

Reading, Mathematics, and Occupational Information

Item Comparisons

Comparison Number	Comparison
1	$White_1$ vs. $Black_1$
2	$White_2$ vs. $Black_2$
3	$White_3$ vs. $Black_3$
4	$White_4$ vs. $Black_4$
5	$White_5$ vs. $Black_5$
6	$White_1$ vs. Oriental
7	$White_1$ vs. Cuban $American_1$
8	$White_2$ vs. Cuban $American_2$
9	$White_3$ vs. Cuban $American_3$
10	$White_4$ vs. Cuban $American_4$
11	$White_5$ vs. Cuban $American_5$
12	$White_1$ vs. American Indian
13	$White_1$ vs. Puerto Rican American
14	$Black_1$ vs. Oriental
15	$Black_1$ vs. Cuban $American_1$
16	$Black_2$ vs. Cuban $American_2$
17	$Black_3$ vs. Cuban $American_3$
18	$Black_4$ vs. Cuban $American_4$
19	$Black_5$ vs. Cuban $American_5$
20	$Black_1$ vs. American Indian
21	$Black_1$ vs. Puerto Rican American
22	Oriental vs. Cuban $American_1$
23	Oriental vs. American Indian
24	Oriental vs. Puerto Rican American
25	Cuban $American_1$ vs. American Indian
26	Cuban $American_1$ vs. Puerto Rican American
27	American Indian vs. Puerto Rican American
28	Rural vs. Urban
29	Male vs. Female

TABLE 2

Percentage of Biased Comparisons Among Ethnic Groups

Reading Total

Comparison	Percentage Biased
White-Black[a]	4.45
White-Oriental	0.00
White-Cuban[b]	6.67
White Indian	0.00
White-Puerto Rican	0.00
Black-Oriental	1.69
Black-Cuban[c]	6.94
Black-Indian	1.67
Black-Puerto Rican	0.00
Oriental-Cuban	3.33
Oriental-Indian	0.00
Oriental-Puerto Rican	0.00
Cuban-Indian	10.00
Cuban-Puerto Rican	1.67
Indian-Puerto Rican	0.00
Rural-Urban	0.00
Male-Female	0.00

[a]Based upon five White-Black comparisons.

[b]Based upon five White-Cuban comparisons.

[c]Based upon five Black-Cuban comparisons.

TABLE 3

Percentage of Biased Comparisons Among Ethnic Groups

Mathematics Total

Comparison	Percentage Biased
White-Black[a]	3.00
White-Oriental	0.00
White-Cuban[b]	0.26
White-Indian	11.84
White-Puerto Rican	0.00
Black-Oriental	4.17
Black-Cuban[c]	1.88
Black-Indian	2.60
Black-Puerto Rican	0.00
Oriental-Cuban	0.00
Oriental-Indian	7.04
Oriental-Puerto Rican	0.00
Cuban-Indian	2.74
Cuban-Puerto Rican	0.00
Indian-Puerto Rican	1.28
Rural-Urban	0.00
Male-Female	0.00

[a]Based upon five White-Black comparisons.

[b]Based upon five White-Cuban comparisons.

[c]Based upon five Black-Cuban comparisons.

TABLE 4

Percentage of Biased Comparisons Among Ethnic Groups

Occupational Information

Comparison	Percentage Biased
White-Black[a]	3.65
White-Oriental	0.00
White-Cuban[b]	0.00
White-Indian	0.00
White-Puerto Rican	0.00
Black-Oriental	2.56
Black-Cuban[c]	4.66
Black-Indian	0.00
Black-Puerto Rican	0.00
Oriental-Cuban	0.00
Oriental-Indian	0.00
Oriental-Puerto Rican	0.00
Cuban-Indian	0.00
Cuban-Puerto Rican	0.00
Indian-Puerto Rican	0.00
Rural-Urban	0.00
Male-Female	2.50

[a]Based upon five White-Black comparisons.

[b]Based upon five White-Cuban comparisons.

[c]Based upon five Black-Cuban comparisons.

Fishbein, Ronald L. Address: Michigan Educational Assessment Program, Box 420, Lansing, Michigan 48902. Title: Research consultant. Degrees: B.A. S.U.N.Y. at Buffalo, M.A. New York University, Ph.D. The Florida State University. Specialization: Educational measurement and evaluation.